End to End ML Pipeline - Part 3

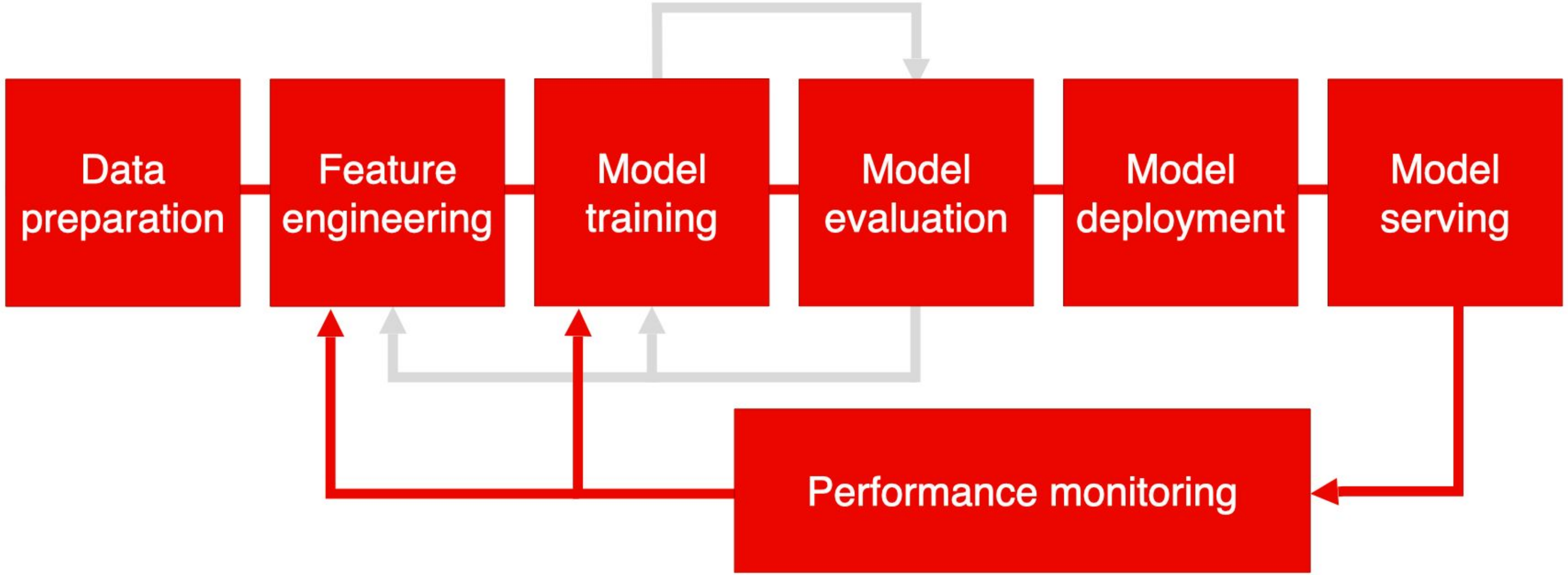
Vithika Karan

ProjectPro

Introduction

Vithika Karan

Model Monitoring



Reference: Evidently AI

ProjectPro

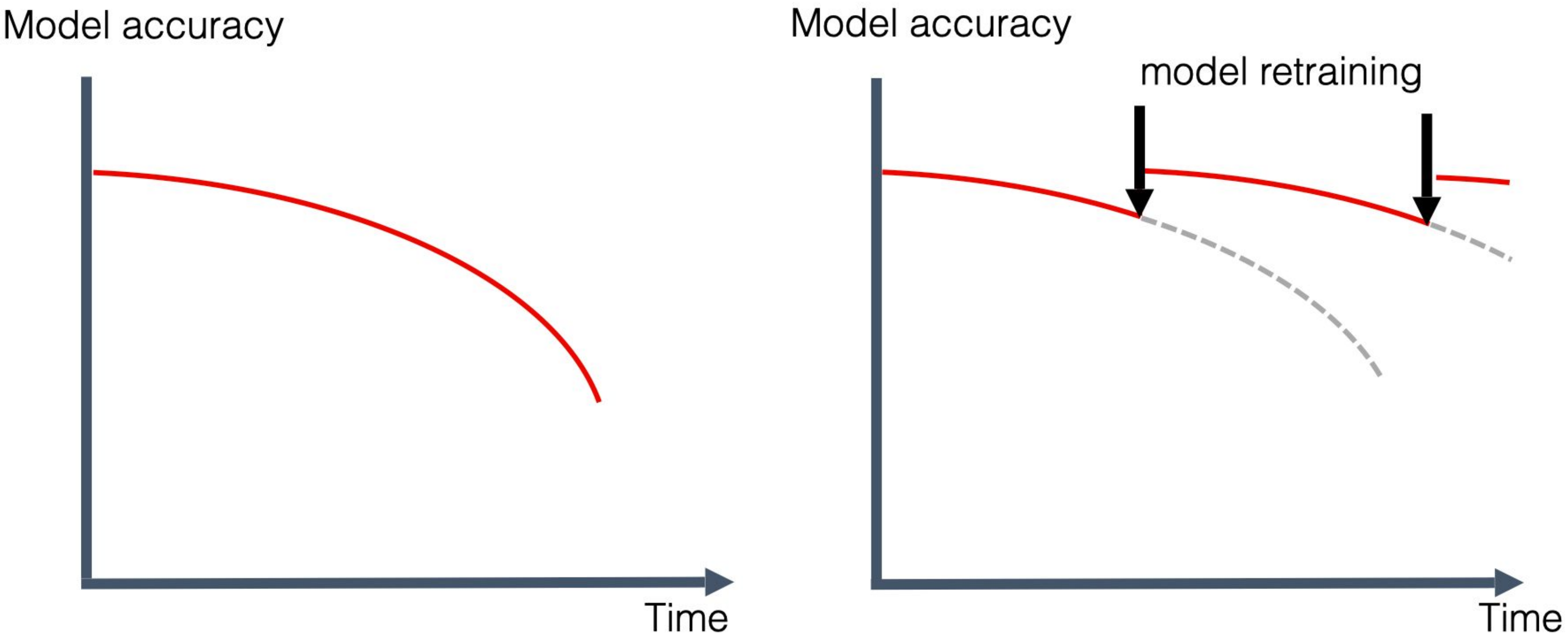
How to do model monitoring?

Data Quality Checks: Ensure data reliability, detect missing values, duplicates, and schema inconsistencies.

Data Drift Detection: Compare new data with historical data to identify distribution and feature changes.

Model Drift Detection: Monitor key metrics for declining model performance, trigger actions like retraining or notifications.

Action Triggers: Set up automated responses on different stages to detected issues for proactive model maintenance.



Reference: Evidently AI

Data Drift

Data Drift:

- Change in statistical properties of input data.
- Occurs in production when data differs from training data.

Importance of Data Drift Monitoring:

- Ensures model remains accurate over time.
- Helps adapt models to changing data patterns.

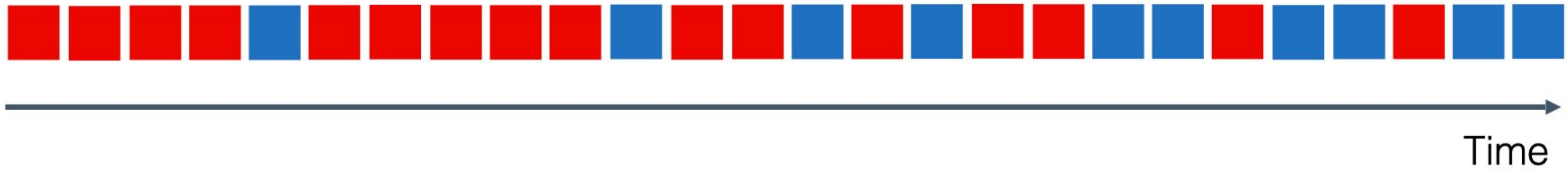
Data Drift vs Concept Drift:

- Data Drift: Shifts in input feature distributions.
- Concept Drift: Shifts in relationships between inputs and outputs.

Data Drift vs Prediction Drift:

- Data Drift: Changes in model input data.
- Prediction Drift: Changes in model outputs.

Incoming data



Feature distribution: sales_channel

Model quality: accuracy over time

Data drift

$P(X)$

Concept drift

$P(Y|X)$

Data drift

$P(X)$

Prediction drift

$P(Y)$

How to detect Drift?

ProjectPro

- Summary Statistics:
  - Compare key stats like mean, median, variance.
  - Example: Check if current mean is within two standard deviations from reference.

- Statistical Tests:
  - Use tests like Kolmogorov-Smirnov for numerical features.
  - Assess if datasets are from different distributions.
  - Output: p-value indicates confidence level.

- Distance Metrics:
  - Measure distance between distributions.
  - Examples: Wasserstein Distance, Jensen-Shannon Divergence.
  - Output: Drift score reflects extent of change.

- https://www.evidentlyai.com/blog/data-drift-detection-large-datasets

Statistical Tests

- Kolmogorov-Smirnov Test (KS Test):
  - Compares two samples to assess if they're from same distribution.
  - Output: p-value, <0.05 suggests different distributions.
  - Used for Numerical Data

- Wasserstein Distance (WD):
  - Measures effort to turn one distribution into another.
  - Output: Absolute value or normed WD, interpretable drift measure.
  - Used for Numerical Data

- Jensen-Shannon Divergence (JS):
  - Calculates difference between probability distributions.
  - Output: Score between 0 and 1, interpretable as similarity.
  - Used for Categorical and Numerical Data

Model Drift and Detection

- Model Drift:
  - Model drift occurs when a machine learning model's performance declines over time due to changes in the underlying data or environment.

- Detecting Model Drift:
  - Compare model performance metrics over time.
  - Monitor key indicators such as accuracy, precision, recall, and F1 score.
  - Look for significant deviations from baseline performance.

- Action Triggers for Model Drift:
  - Set up automated triggers to respond to detected model drift.
  - Actions may include retraining the model, updating training data, or adjusting model parameters.

- Example of Model Drift:
  - Illustrate with a scenario: A spam detection model that gradually becomes less effective over time as spammers adapt their tactics.

Thank you